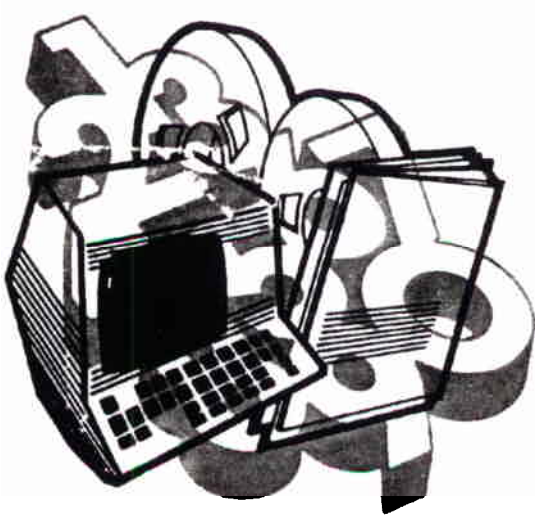
Bureau of the Census

1993 Annual Research Conference

March 21–24, 1993

Doubletree Hotel – National Airport
300 Army Navy Drive
Arlington, Virginia 22202

Proceedings



U.S. Department of Commerce
Economics and Statistics Administration
BUREAU OF THE CENSUS

MULTIPLE IMPUTATION OF MISSING DATA IN NHANES III

Joseph L. Schafer, Pennsylvania State University,
Meena Khare, National Center for Health Statistics, and
Trena M. Ezzati-Rice[1], National Center for Health Statistics.

ABSTRACT

The National Health and Nutrition Examination Survey (NHANES) collects important nutritional and health-related data on the civilian noninstitutionalized U.S. population and important subgroups. High rates of unit nonresponse in NHANES, together with some residual item nonresponse, lead to high rates of missingness on key survey variables. We discuss a project to statistically adjust for nonresponse in NHANES III, Phase 1 (1988-1991) using multiple imputation (Rubin 1987). A data file consisting of 27 key variables for 12,392 sampled individuals was multiply imputed for both item and unit nonresponse, using techniques of iterative Bayesian simulation via Markov chains described by Schafer (1991, 1993). The processes of data editing, model selection, and simulation of missing data are discussed along with related computational issues. Exploratory analysis of the imputed values suggests that both the marginal distributions of the survey variables, and important relationships between them, were accurately preserved. Multiple-imputation interval estimates for scalar quantities of interest (means, proportions, subdomain means, etc.) were, in some cases, dramatically wider than corresponding intervals that ignored the missing-data uncertainty. This project is significant in that it represents the first successful implementation of proper multiple-imputation methodology in a large multivariate setting, and consequently, it gives useful insight into the feasibility of multiply imputing NHANES and other large multipurpose sample surveys on an ongoing basis.

KEYWORDS

Bayesian Simulation, Gibbs Sampling, Sample Surveys

1. INTRODUCTION

1.1 OVERVIEW OF NHANES III

The National Health and Nutrition Examination Survey (NHANES) is a periodic national survey conducted by the National Center for Health Statistics (NCHS). The ongoing Third National Health and Nutrition Examination Survey (NHANES III) is the seventh in a series of similar surveys conducted by NCHS since the 1960's. NHANES III is designed to provide national statistics on health and nutritional status for the civilian noninstitutionalized population aged 2 months and older. The data are obtained through household interviews, which collect socio-demographic and medical history information, and through standardized physical examinations, which include a variety of physical measurements and physiological tests. The physical examinations are conducted in specially equipped mobile examination centers (MEC's) that are transported to each survey location. NHANES III has been divided into two 3-year surveys--Phase 1 (1988-1990) and Phase 2 (1991-93)--in order to provide national estimates for each 3-year period as well as for all six years.

---

[1]Views expressed are those of the authors and do not necessarily reflect those of the National Center for Health Statistics. Thanks to Donald B. Rubin and Roderick J. A. Little for valuable suggestions throughout this project including an extensive written report (Little and Rubin 1992)

NHANES III is based on a complex, multistage area probability sample design with a total sample size of approximately 40,000 persons. Children under 5 years of age, adults age 60 years and older, Mexican Americans, and Black Americans are sampled at higher rates than other persons in order to provide accurate estimates within important demographic subpopulations. Details of the sample design of NHANES III have been previously published by Ezzati et al. (1992). Phase 1, with which this article is concerned, had a total sample size of 20,277 persons.

1.2 NONRESPONSE IN NHANES III

NHANES, like most sample surveys, suffers from both unit and item nonresponse. In an attempt to maximize response rates, special procedures have been implemented including extensive publicity in each survey location, a home examination especially targeted for the older population, a remuneration to all survey participants, and a report of major medical findings. Despite these efforts, however, nonresponse rates remain quite high, both in the personal interview and in the MEC examination. Experience with previous NHANES surveys has led us to expect overall unit nonresponse rates of about 10% and 25% in the personal interview and physical examination, respectively. These high rates of unit nonresponse, when combined with various levels of item nonresponse pertaining to specific questions, examination components, or physical measurements, lead to rather high overall rates of missingness on some key survey variables.

The patterns of nonresponse in NHANES III are heavily influenced by the process of data collection, which occurred primarily in three stages:

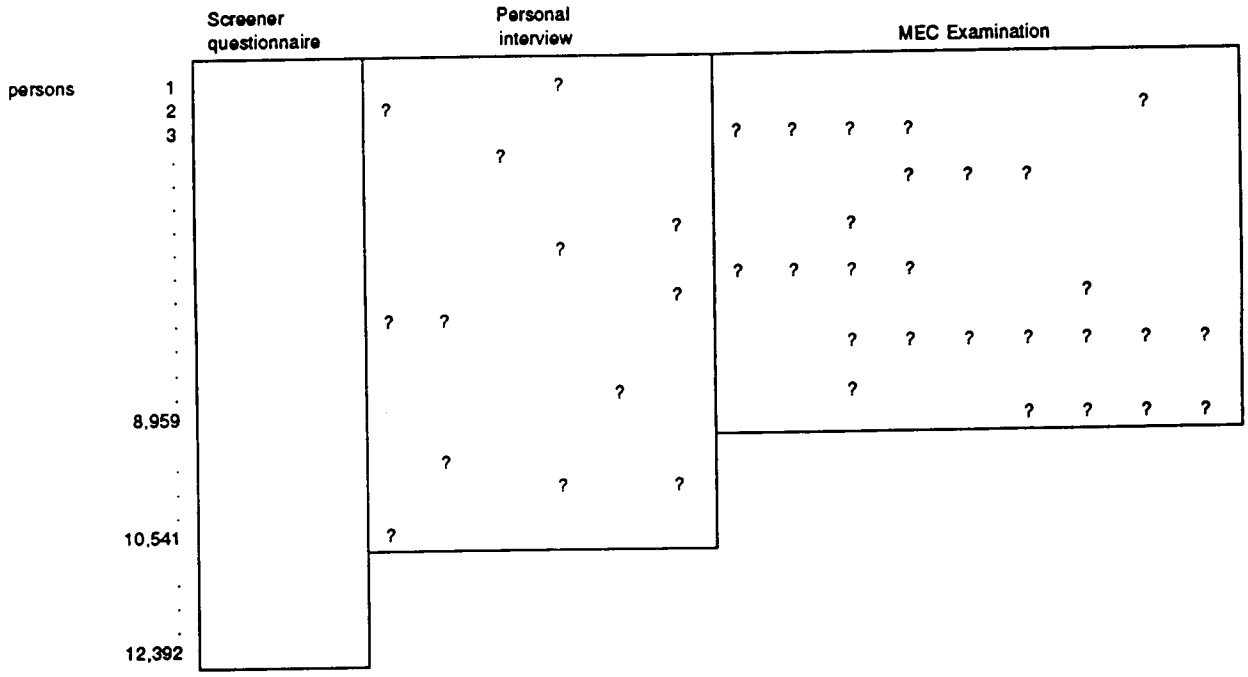
1. *Household screening.* When a household was selected into the NHANES III sample, a brief screening interview was conducted to determine household size and the age, sex, and race of every household member. This information was required for the final stage of sampling in which individuals were selected within households. As a byproduct of this screening procedure, the basic demographic characteristics--age, sex, and race--are known for each sampled person; no data are missing for these items.

2. *Personal interview.* After the household screening and final stage of sampling, NHANES personnel conducted interviews to obtain detailed health and nutritional information for sampled persons. Among the 20,277 sampled persons in Phase 1 of NHANES III, 17,464 (86%) were successfully interviewed. Refusal or inability to answer specific interview questions led to some additional item nonresponse, at typical rates of about 1-5% per item.

3. *MEC examination.* Upon completion of the personal interview, sampled persons were requested to report to the MEC for the physical examination. Among the 17,464 interviewed persons in Phase 1, 15,884 (91%) reported to the subsequent physical examination, resulting in an overall examination rate of 78%. The MEC examination included a number of components or groups of procedures: anthropometric (body size) measurements, blood pressure readings, blood and urine specimens, 24-hour dietary recall, dental examination, spirometry, etc. In the examinations, of course, not all items were successfully recorded for all examinees. Occasionally, examinees left the MEC before the examination could be completed, or refused a specific procedure (e.g., venipuncture), causing an entire group of items to be missing. Data recording errors and other mistakes by personnel also caused single items or groups of items to be missing. Among examined persons, nonresponse rates for individual MEC items were on the order of 5-8%.

At the end of this data collection process, many key variables from the MEC examination were missing at rates relative to the entire sample of 30% or more. A schematic representation of data from NHANES III Phase 1 depicting the pattern of missingness is shown in Figure 1.

FIGURE 1. Schematic representation of data from NHANES III Phase 1 showing unit and item nonresponse, sampled adults only. Rows represent sampled persons, columns represent survey variables, and question marks (?) denote missing items.



1.3 NONRESPONSE ADJUSTMENTS

As described above, the major portion of the missing data in NHANES is due to unit nonresponse in the interview and/or examination, with item nonresponse comprising a relatively small part. It is common practice to compensate for unit nonresponse by weighting-class adjustments (Madow, Olkin, and Rubin, 1983; Cox, 1991). Weighting-class adjustments are usually performed by grouping respondents and nonrespondents together into a relatively small number of classes or cells, assigning the nonrespondents survey weights of zero, and inflating the weights of the remaining respondents proportionately so that the total weight of the units within cells is preserved. After this weighting adjustment, any residual item nonresponse that remains is typically handled by imputation--i.e., the missing items are filled in with plausible values gleaned from other similar units in the dataset, or with predicted values obtained from a model.

Weighting-class adjustments for unit nonresponse have been used in previous NHANES surveys and were also planned for NHANES III (Ezzati and Khare, 1991, 1992). Survey weights for the examined persons were inflated up to the level of the full sample. This adjustment was performed within broad classes defined by geography, demographic variables, and family income. Since income is available only from the personal interview and is itself subject to nonresponse, it had to be imputed before the weighting-class adjustment could be carried out. After the weighting adjustment, little or no imputation was used to compensate for the remaining item nonresponse in MEC examination; missing items were typically left blank and omitted from further analyses.

One disadvantage of this weighting-class approach is that very little information obtained from the personal interview was used in the nonresponse adjustment for the non-examined. In Phase 1, over one-third (36%) of the non-examined persons had been successfully interviewed; with the exception of income, however, none of the information gleaned from these interviews was used in the formation of weighting classes.

Yet, the personal interview provided many variables that are potentially powerful predictors for some of the MEC examination items. The two most striking examples of this are:

1. *Body measurements.* In the personal interview, sampled persons were asked to report their own height and weight. One would naturally expect these self-reported values to be highly correlated with many of the body measurements recorded in the subsequent MEC examination, such as height, weight, waist and buttocks circumference, skin fold measurements, etc. Exploratory analyses revealed that self-reported height and weight were indeed very highly correlated with MEC height and weight ($r = .876$ and .967 on the log scale, respectively).

2. *Blood pressure.* As part of the interview, blood pressure readings were taken in the subject's home. These blood pressure readings from the interview were found to be highly correlated with blood pressure readings from the subsequent MEC examination, with observed correlations on the order of $r = .6$.

In addition, many other responses to interview questions regarding smoking, hypertension and high blood cholesterol, etc. are potentially useful predictors for many MEC items. Thus, there seem to be significant potential gains, both in reducing nonresponse bias and increasing precision, from including more of the interview variables in the nonresponse adjustment.

1.4 EXAMINING IMPUTATION ALTERNATIVES

In 1992, NCHS initiated a project to investigate alternatives to the current NHANES nonresponse adjustment methodology, including imputation (Little and Rubin, 1992). Imputation, although typically more difficult to carry out in practice than weighting-class adjustments, offers some potentially important advantages including the reduction of variance and the opportunity to use more covariate information (Little, 1986). Moreover, through the technique of multiple imputation (Rubin, 1987), it is possible to assess the impact of missing-data uncertainty on the variances of estimators and revise variance estimates to reflect this additional uncertainty.

Applications of multiple imputation to large surveys such as NHANES have been previously hampered by the difficulty of generating proper multiple imputations in multivariate settings. In multivariate datasets, complex patterns of missingness cause the predictive distributions of the missing values, even under simple probability models, to be intractable and difficult to simulate directly. Recent advances in techniques of Bayesian computation, however, now make it possible to generate proper multiple imputations in multivariate settings under a variety of useful models for both continuous and categorical data (Schafer, 1991). Multiple imputations can now be routinely generated using iterative simulation schemes based on Markov chains, including the Gibbs sampler and the Metropolis algorithm (Geman and Geman, 1984; Gelfand and Smith, 1990; Müller, 1991).

1.5 THE NHANES III MULTIPLE IMPUTATION PROJECT

To test the applicability of this new multiple-imputation methodology to NHANES, a data file was prepared consisting of approximately 30 key variables from the screener questionnaire, personal interview, and MEC examination in Phase 1 of NHANES III. Important features of this dataset, including patterns and rates of nonresponse, are discussed in Section 2. In Section 3 we describe the process by which we edited the data and devised a multivariate model to describe the joint probability distribution of all the variables for purposes of imputation. Section 4 discusses the computational details of estimating the parameters of this multivariate model and simulating the multiple imputations. Exploratory analyses of the imputed datasets, including graphical displays and multiply-imputed interval estimates, are described in Section 5, and Section 6 presents concluding discussion and final remarks.

TABLE 1. MEC variables in the NHANES III imputation project.

Name	% missing	Description
Body measurements		
HT	31.5	height
WT	33.2	weight
WST	32.9	waist circumference
BUT	32.8	buttocks circumference
Blood pressure		
BP1K1D	33.2	first systolic pressure
BP1K5D	33.3	first diastolic pressure
BP2K1D	33.4	second systolic pressure
BP2K5D	33.5	second diastolic pressure
BP3K1D	33.4	third systolic pressure
PB3K5D	33.6	third diastolic pressure
Lipids		
TCRES	33.3	total serum cholesterol
HDRES	33.9	HDL cholesterol

2. THE DATA

2.1 MEC VARIABLES

Because the number of variables recorded in NHANES is enormous, we decided to narrow the goal of our project to producing a set of good quality multiple imputations for just a few key variables from the MEC examination. In particular, we decided to focus attention on just twelve variables from three MEC components--body measurements, blood pressure, and lipids. Also, because the personal interview and examination procedures were substantially different for adults and children, we restricted our study to the 12,391 adults (age 17 years and older) in the Phase 1 sample. Among these adults, only 8,959 (72.5%) completed both the interview and the examination; 1,851 (14.9%) were neither interviewed nor examined, and 1,582 (12.8%) were interviewed but not examined. Following the pattern of Figure 1, every adult who missed the interview also missed the examination. The twelve MEC variables with their overall rates of missingness, reflecting both unit and item nonresponse, are listed in Table 1.

2.2 PRE-MEC VARIABLES

Although our primary interest was in imputing the twelve MEC variables, we also included in our analysis a number of additional variables from the screening and personal interviews. These pre-MEC variables were judged to contain potentially valuable information for imputing the missing MEC items. Pre-MEC variables were explicitly modeled together with the MEC variables, and missing MEC variables were imputed conditionally upon these pre-MEC variables whenever available. When the pre-MEC variables were missing, they too were imputed conditionally upon any pre-MEC or MEC variables that were present.

When imputing variables subject to nonresponse, conditioning on auxiliary variables has some well known benefits (e.g., Little, 1986). First, if the probability of nonresponse for a variable in question is related to the auxiliary variables, then conditioning will tend to reduce nonresponse bias. Second, if the auxiliary variables are related to the variable in question, then conditioning will also tend to reduce variance, in much

the same way that the classical survey techniques of ratio and regression estimation tend to reduce variance. For the purpose of reducing mean squared error of prediction, then, an imputation procedure ought to make maximal use of whatever covariate information is available.

Another important, but less well known, benefit of including auxiliary variables arises when attempting to reflect missing-data uncertainty. Multiple imputation will provide valid inferences only if the imputations exhibit enough variability to represent our true state of knowledge, in a conditional or *a posteriori* Bayesian sense, about the missing values (Rubin, 1987). Omitting an auxiliary variable from the imputation procedure is equivalent to specifying, say, a regression model in which the coefficient of the auxiliary variable is set to zero *a priori*. Fixing parameters of the imputation model to zero, when the data do not provide strong evidence that they are truly zero, will tend to produce multiple imputations having too little variability.

One set of auxiliary variables that requires careful consideration is the set that conveys information about the sample design. Surveys with complex sampling plans have important features--unequal probabilities of selection, stratification, and clustering--that distinguish them from simple random samples. The observational units in complex surveys are typically not exchangeable and cannot be appropriately described by simple probability models that assume, for example, that units are independent and identically distributed. Recently, multiple-imputation inference has been criticized for failing to produce accurate variance estimates in some hypothetical counterexamples (Fay, 1992). In all of these "counterexamples," however, the multiple imputations are not drawn from the correct predictive distribution, the distribution that conditions fully on the observed data including indicators of the sample design; invariably, some important information is left out of the imputation model. In order to guarantee that multiple imputation inferences are valid, essential information about the sample design must be included in the analysis.

TABLE 2. Pre-MEC auxiliary variables in the NHANES III imputation project.

Name	% missing	Description
PSU identifier		
STAND	0.0	examination location (101-144)
Demographics		
AGE	0.0	age (17-39, 40-59, 60+)
SEX	0.0	gender (male, female)
RACE	0.0	race/ethnicity (Black, Mex-Amer, Other)
Personal interview		
ACTV	20.6	self-reported activity status
AD1	18.5	diabetes ever diagnosed?
AE2	19.3	hypertension ever diagnosed?
AE7	62.4	high cholesterol ever diagnosed?
AF10	20.6	heart attachk ever diagnosed?
AR3	18.3	smoke cigarettes now?
ALCO	18.6	beer/wine/liquor?
AHT	22.7	self-reported height
AWT	21.6	self-reported weight
ASYS	21.9	interview average systolic b.p.
ADIAS	21.9	interview average diastolic b.p.

Finally, apart from considerations of mean squared error and variance estimation, we also felt it was essential to include auxiliary variables to preserve important statistical relationships in the dataset, especially those relationships that may be of interest to potential secondary users of the data. In large

datasets like NHANES, relationships that seem trivial in other settings may be of great scientific interest. For example, our exploratory analyses revealed a highly significant ($p < .0001$) relationship between response to the interview question, "Has a doctor ever told you that you had a heart attack?" and HDL cholesterol; the partial correlation between these two variables given other important covariates, however, was only .078 (see Section 3.2 and Table 5). If this interview question was left out of the imputation procedure, then the mean squared error for estimating population average levels of HDL cholesterol would increase by only a negligible amount. If the imputed data were later used in a secondary analysis of relationships between heart disease and lipids, however, then leaving the question out of the imputation model could seriously dampen this small but real relationship.

All of the above arguments tend to favor large imputation models over small ones, encouraging us to impute conditionally on any auxiliary variable that might be important. Balanced against these considerations, of course, are computational limitations that may prevent us from fitting a model as large as we would like. As the number of variables grows, memory requirements for the model-fitting and imputation algorithms increase dramatically. Keeping the model to a reasonable size required us to make some tough choices in variable selection. Starting with about twenty candidate auxiliary variables, we reduced the list, on the basis of exploratory regression analyses (described below) and {\it a priori} considerations, to the fifteen variables shown in Table 2. Information on the stratified cluster design of NHANES was reflected in STAND, a 44-level categorical variable indicating the mobile examination location or primary sampling unit (PSU) to which a person belonged. Further information pertinent to the final stage of sampling was contained in the demographic variables AGE, SEX, and RACE. Eleven variables from the personal interview were included because they were found to have statistically significant and scientifically important relationships to one or more of the twelve MEC items.

3. BUILDING THE MODEL

3.1 DATA CLEANING AND REMOVAL OF OUTLIERS

Initial exploration of the NHANES III imputation data revealed a substantial number of outliers, particularly in the body measurements. These outliers for the most part reflected gross errors in the data recording and capture process; if allowed to remain in the dataset, they could have exerted an undue influence on the parameter estimates and artificially inflated the variability of the multiple imputations. For these reasons, the data were screened by a variety of informal techniques, and observations that were identified as being clearly erroneous were deleted (recoded as missing).

First, a number of measurements were deleted simply because they were out of the range of physical plausibility--e.g., self-reported or measured heights less than 80 cm or diastolic blood pressure readings below 20 mm. After this initial variable-by-variable screening for out-of-range values, additional unusual observations were identified using bivariate scatterplots. Scatterplots were created for pairs of body measurements that are known to be highly correlated. Points located far outside the bivariate point clouds were highlighted and identified interactively using an X11 graphics window on a Unix workstation and the "identify" function in S (Becker, Chambers, and Wilks, 1988). These questionable data were not automatically deleted but were earmarked for further study.

Finally, regression models were fit to each of the body measurements using the other body measurements as predictors (suspect points were excluded from the fitting). Each questionable body measurement was then compared to its fitted value from the regression, and if the standardized residual was exceedingly large, the offending measurement was deleted. The decision to delete a suspect measurement was carried out not according to a set of strict rules, but on an informal, case-by-case basis. The approach used was conservative in that if any doubt existed about whether a recorded value was erroneous, the value was allowed to remain. It is known that gross errors occur naturally in self-reported height and weight--e.g., some obese persons will seriously underreport their actual weights. Every reasonable effort

was made to preserve the integrity of the original data. Based on this informal analysis, a total of about 150 measurements were deleted.

3.2 REGRESSION MODELING TO CHOOSE AUXILIARY VARIABLES

After removing the outliers, we performed some exploratory regression analyses to investigate how the pre-MEC auxiliary variables under consideration were related to the MEC variables of primary interest. Linear regression models were fit to the following eight dependent variables from the MEC examination: LGHT, LGWT, LGWST, and LGBUT (height, weight, waist circumference and buttocks circumference on a log scale), LGDSYS (the average of the three systolic blood pressures BP1K1D, BP2K1D, BP3K1D on a log scale), DDIAS (the average of the three diastolic blood pressures BP1K5D, BP2K5D, BP3K5D), LGTCRES (log serum total cholesterol), and LGHDRES (log HDL cholesterol). The goal of this modeling effort was to identify important relationships between the auxiliary variables and the MEC variables, and to reduce the twenty or so candidate auxiliary variables to a somewhat smaller set to be included in the final multivariate imputation model.

The philosophy of selecting variables for an imputation model should be somewhat different from the traditional variable-selection strategies found in textbooks on multiple regression. In the traditional approach, a variable is not included in a model unless it is deemed "significant"--i.e., unless one cannot reject the hypothesis that its coefficient is zero at some prespecified level such as .05. This traditional approach places a high priority on model *parsimony* (i.e., having no unnecessary covariates) and on model *interpretability*. In imputation, however, the primary goal is *prediction*-- to generate imputed values with the desirable statistical properties outlined in Section 2.2. When building an imputation model, then, one ought also to include variables that fall into these categories:

1. Variables that are considered important on *a priori* grounds. For example, since we knew that many of the results of NHANES would be published for subdomains defined by age, sex, and race, we considered it essential for the full AGE×SEX×RACE effect to be included in the imputation model.

2. Variables that conveyed essential information about the complex sample design.

3. Variables that have large coefficients and large standard errors, even if the coefficients may not be significantly different from zero. Including these variables may be important for creating multiple imputations with enough variability to reflect the actual uncertainty about the missing values.

To simplify the modeling effort at this stage, we fit regressions using only the complete cases--the sampled adults for whom all the MEC and auxiliary variables were observed--which numbered about 2,900. The regression model for predicting each of the eight dependent MEC variables included the other seven as predictors--e.g., the model for LGTCRES included LGDSYS, DDIAS, LGHDRES, LGHT, LGWT, LGWST, and LGBUT--as well as various combinations of candidate auxiliary variables. All regression models included main effects for STAND, a 43 degree-of-freedom set of dummy indicators to distinguish between primary sampling units, and the full 17 degree-of-freedom cross-classification by AGE, SEX, and RACE. All other pre-MEC variables were included either as continuous or as single degree-of-freedom dummy indicators. Self-reported height and weight, as well as systolic blood pressure from the interview, were expressed on a log scale.

Including about 100 regressors left us with 2,800 degrees of freedom for assessing significance. With such a large sample size, even very small effects in the data could be detected with high power. As a result, virtually every pre-MEC variable was found to be significantly related to at least one of the MEC variables. The criterion of statistical significance alone would have included almost every variable and would have produced a model of unmanageable size. In the end, we had to eliminate a few auxiliary variables that did not seem important on *a priori* grounds, and whose effects, although statistically significant, were of relatively small magnitude in comparison to the others.

Some results of the "final" regression models, which include only those auxiliary variables chosen for the imputation model, are displayed in Tables 3-5. Table 3 displays the eight values of the multiple correlation R^2, which provide insight into how well the missing MEC items can be predicted on the basis of other MEC and pre-MEC variables. The matrix in Table 4 displays levels of significance for each of the pre-MEC variables in predicting the eight dependent MEC variables. In Table 4, as well as in Table 5, DEMOG refers to the 17 degrees-of-freedom effect of AGE×SEX×RACE.

One useful measure of effect size in multiple regression is the {\it partial correlation coefficient}, which measures the correlation between one explanatory variable and the dependent variable given all the other explanatory variables. The partial correlation is easily calculated as

$$r = \sqrt{\frac{F}{F + df_2/df_1}}$$

where F is the F-statistic for testing the null hypothesis that the explanatory variable of interest has no effect, and df_1 and df_2 are the numerator and denominator degrees of freedom, respectively. When $df_1 = 1$, the partial correlation is typically given the same sign as the estimated coefficient of the explanatory

TABLE 3. Multiple correlation R^2 statisitcs for prediction of MEC variables in exploratory regression models.

Dependent Variable

LGHT	LGWT	LGWST	LGBUT	LGDSYS	DDIAS	LGTCRES	LGHDRES
.923	.976	.876	.882	.712	.605	.319	.318

TABLE 4. Significance of pre-MEC variables in exploratory regression analyses: *= significant the level .10 **= significant at the level .05; *=significant at the level .01.**

Dependent Variable

	LGHT	LGWT	LGWST	LGBUT	LGDSYS	DDIAS	LGTCRES	LGHDRES
STAND	**	***	***	***	***	***	**	***
DEMOG	***	***	***	***	***	***	***	***
ACTV	*	**	***	***				
AD1		*	***	***	***	***	***	
AE2					***	***	***	
AE7	**			**			***	***
AF10		*			*			***
AR3		***	***	***		**		***
ALCO	**	**		*	**			***
LGAHT	***				***			***
LGAWT		***	*				***	
LGASYS					***	***		*
ADIAS					***	***		

variable of interest. When $df_1 > 1$, as in STAND and DEMOG, r can be interpreted as the partial correlation between the dependent variable and the best linear combination of the components of the explanatory variable. Partial correlation effect sizes between the MEC and pre-MEC variables are displayed in Table 5. For interpretation, we note that with df_2=2,800 error degrees of freedom, a partial correlation of only r=.037 is significant at the .05 level when df_1=1. For df_1=17 and df_1=43, significance at the .05 level is achieved by partial correlations of .099 and .144, respectively.

3.3 THE IMPUTATION MODEL

The most straightforward way to generate proper multiple imputations in a multivariate setting is to specify a parametric probability model for the complete (missing and observed) data along with a prior distribution for the parameters, and then simulate values from the conditional distribution of the missing data given the observed data. Because the NHANES III imputation file contains both continuous and categorical variables, we chose to work with a special case of the model for mixed continuous and categorical multivariate data introduced for discriminant analysis by Krzanowski (1980, 1982) and applied to incomplete multivariate data by Little and Schlucter (1985).

TABLE 5. Partial correlation effect sizes for pre-MEC variables in exploratory regression analyses. Except for STAND and DEMOG, all pre-MEC variables are one degree of freedom with signs ( ) indicating the direction of the effect.

	Dependent Variable							
	LGHT	LGWT	LGWST	LGBUT	LGDSYS	DDIAS	LGTCRES	LGHDRES
STAND	.15	.20	.20	.18	.24	.28	.15	.16
DEMOG	.15	.26	.45	.53	.21	20	.22	.19
ACTV	.04	-.04	.08	.06	-.01	-.01	.00	.00
AD1	-.02	-.03	-.06	.10	-.07	.09	.07	.00
AE2	-.01	.02	.01	-.03	.08	.06	-.06	-.01
AE7	-.05	.02	.02	.05	.00	.00	-.35	.06
AF10	.02	-.03	.01	.00	-.03	.03	.00	.08
AR3	-.01	.07	-.10	.07	.02	.04	-.01	.08
ALCO	-.04	-.04	.02	.03	.05	-.02	.00	-.15
LGAHT	.84	.00	.00	-.02	.05	-.02	-.01	.08
LGAWT	.02	.74	.03	.02	-.02	-.03	-.06	-.02
LGASYS	.02	-.01	.03	-.01	.57	-.25	.01	-.03
ADIAS	-.02	.00	.00	-.01	-.25	.53	.00	.02

Let Y denote the matrix of complete data, which can be partitioned as Y=(W,Z), where W is an n×p matrix of categorical variables, and Z is an n×q matrix of continuous variables. Let $W_1,W_2,\ldots,W_p$ and $Z_1,Z_2,\ldots,Z_q$ denote the variables in W and Z, respectively. Suppose that the categorical variable W_j takes d_j possible levels, so that each row of W can be classified into a cell of a p-dimensional contingency table with total number of cells equal to $D=\prod_{j=1}^{p} d_j$. Let $\{x_{ijk\cdots t}\}$ denote the cell counts of this contingency table, where $x_{ijk\cdots t}$ is the number of rows of W for which $W_1 = i, W_2 = j, \cdots, W_p = t$. It will also be notationally convenient to index the cells of the contingency table by the single subscript d, ranging from 1 to D, so that the cell frequencies may be written more simply as $\{x_d\}$.

The multivariate distribution for Y is most easily described in terms of the marginal distribution of W and the conditional distribution of Z given W. We assume that the marginal distribution of W is a multinomial distribution on the cell counts $\{x_{ijk\cdots t}\}$, with cell probabilities denoted by $\pi = \{\pi_{ijk\cdots t}\} = k\{\pi_d\}$. Conditionally upon W, we assume that the rows of Z are multivariate normal with mean vectors that vary between cells of the contingency table, but with a common covariance structure for all cells. That is, given that an individual's categorical variables determine that he or she should be placed into cell d, then his or her values of $(Z_1,Z_2,\cdots,Z_q)$ are assumed to be $N(\mu_d,\Sigma)$ independently of all other individuals. Letting $\mu = (\mu_1,\mu_2,\cdots,\mu_D)^T$ denote the D×q matrix of conditional means, we can write the unknown parameters as $\theta = (\pi,\mu,\Sigma)$.

Without any restrictions on θ except the obvious one that $\sum_{d=1}^{D} \pi_d = 1$, this model has $(D-1)+Dq+q(q+1)/2$ free parameters. As the number p of categorical variables grows, the contingency table typically becomes

too sparse to estimate the probabilities π_d for the individual cells, much less the mean vectors μ_d within cells. For this reason, we reduce the dimensionality of the parameter by allowing loglinear constraints on the cell probabilities π, and/or ANOVA-like constraints on the cell means μ. Loglinear constraints are a well known device for fitting parsimonious models to contingency tables (e.g., Bishop Fienberg, and Holland, 1975) and will not be described here. Let A be a D×r design matrix that relates the within-cell means μ to an r×q matrix of regression coefficients β in the manner $\mu=A\beta$, where $\text{rank}(A)=r\leq D$. In other words, we will still allow the means μ_d to vary from cell to cell, but require that each column of μ lie in the r-dimensional linear subspace of R^D spanned by the columns of A. By saturating the loglinear model for π and taking A=I (the identity matrix), we obtain the most general model with (D-1)+Dq+q(q+1)/2 free parameters as a special case.

Among the 27 variables listed in Tables 1 and 2, sixteen (all of the MEC variables in Table 1, plus AHT, AWT, ASYS, and ADIAS) are continuous while the remaining eleven consist of ordered or unordered categories. STAND and RACE are unordered with 44 and 3 levels, respectively; AGE and ACTV are three-point ordinal scales; and the remaining seven variables are dichotomous. Attempts to fit a model with eleven categorical variables proved futile, because the contingency table with $D = 44 \times 3^3 \times 2^7 = 152{,}064$ cells was much too sparse to allow for stable estimation of the within-cell means μ, unless undesirably strong restrictions were introduced on μ through the design matrix A. Further elimination of categorical variables to reduce the dimensionality of the contingency table was undesirable, because we considered all eleven to be important. In particular, retention of the 44-level classification by STAND, even though this variable was one of the main causes of sparseness, was considered essential to ensure that sample-design information was properly reflected in the multiple imputations.

After considering several alternatives, we finally decided to retain only four variables—AGE, SEX, RACE, and STAND—in the categorical portion of the model, treating the other 23 variables as continuous and conditionally multivariate normal given these four. The contingency table for AGE × SEX × RACE × STAND had 792 cells for 12,392 observations. Because of the sample design, this table was filled in rather nicely, with only 157 empty cells. Modeling the six dichotomous variables AD1, AE2, AE7, AF10, AR3, and ALCO, and the three-point ordinal variable ACTV, as continuous and conditionally normal was only a very rough approximation at best. We considered this approximation to be acceptable, however, because these seven variables were not among the variables of primary interest in our study. The variables of greatest interest were the twelve MEC variables listed in Table 1. Pre-MEC variables were intended to serve primarily as predictors, although they too were imputed whenever missing. Moreover, some limited evidence suggests that erroneously modeling the seven discrete variables as continuous did not have a strong adverse effect on the final imputations; when the continuous imputes for these seven were rounded off to the nearest categories, the distributions of the imputed values were quite reasonable and looked very similar to the distributions actually observed in the sample (Section 5.1).

In the final analysis, we modeled the 635 nonempty cells of the contingency table for AGE × SEX × RACE × STAND by a saturated multinomial distribution, treating the 157 empty cells as structural zeros. (This specification had no effect on the distribution of imputed values, because these four variables were never missing.) The 23 remaining "continuous" variables were then modeled as a multivariate normal linear regression. To make the normality assumption more plausible, body measurements, lipids, and systolic blood pressures were expressed on a log scale. Each of the 23 individual regressions included an intercept, 17 dummy indicators to represent the full AGE × SEX × RACE interaction, and 43 dummy indicators to represent STAND, for a total of $23(1 + 17 + 43) = 1403$ estimated regression coefficients and $23(24)/2 = 276$ residual variances and covariances. The total number of unknown free parameters in this model was thus $(635 - 1) + 1403 + 276 = 2313$.

4. MODEL FITTING AND IMPUTATION

4.1 TECHNIQUES FOR ESTIMATION AND SIMULATION

A full description of the techniques we used to fit our model and generate multiple imputations is beyond the scope of this article and can be found in Schafer (1991, 1993); we present only the basic strategy and a few computational details.

An iterative algorithm for maximum-likelihood (ML) estimation with incomplete data under the multivariate model described above is given by Little and Schlucter (1985), and is an example of ECM--"Expectation-Conditional Maximization" (Meng and Rubin, 1992), a generalization of the well known EM algorithm. A general version of ECM for this multivariate model, along with additional algorithms for parameter simulation and multiple imputation described below, have been implemented for use in the statistical package S with external Fortran subroutines (Schafer 1991, 1993).

To simulate missing data under an assumed value of the parameter, such as $\theta = \hat{\theta}$ (the ML estimate), would be relatively straightforward. Under our model, the vector of missing observations for each person has, given his or her observed data, a multivariate normal distribution with parameters that can be calculated by applying a suitable transformation to θ. Multiple simulated versions of the missing data under $\theta = \hat{\theta}$, however, would not be proper multiple imputations. Proper multiple imputations must reflect the uncertainty associated with the fact that θ is not known but merely estimated. Proper multiple imputations can be most easily conceptualized as repeated draws from a Bayesian posterior predictive distribution for the missing data given the observed data. Let Y_{obs} denote the observed data and Y_{mis} the missing data. The posterior predictive density of Y_{mis} given Y_{obs}, or $P(Y_{mis}|Y_{obs})$, is

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs},\theta)\, P(\theta|Y_{obs})\, d\theta, \qquad (1)$$

where $P(\theta|Y_{obs})$ is the posterior density of the parameters given the observed data; in other words, the posterior predictive distribution of the missing data is the conditional distribution of the missing data given the observed data under an assumed θ, averaged over the posterior distribution of θ.

It is important to note that the distribution (1) is the appropriate source of multiple imputations only under the assumption that the nonresponse mechanism is *ignorable*, or that the missing data are *missing at random*, in the sense defined by Rubin (1976, 1987). Despite its name, missing at random does not imply that the missing values are necessarily a simple random sample of all data values. The latter condition is known as *missing completely at random*, which is only a special case of missing at random. Missing at random requires only that the missing values be like a random sample of all values within subclasses defined by observed data. In other words, missing at random does allow the probability that a data value is missing to depend on the value itself, but only indirectly through quantities that are actually observed. Throughout this analysis, we assume that the missing at random assumption holds.

Except in special cases, $P(Y_{mis}|Y_{obs})$ tends to have an intractable form, and direct simulation of Y_{mis} from $P(Y_{mis}|Y_{obs})$ can be prohibitively difficult. It is sometimes possible, however, to simulate $P(Y_{mis}|Y_{obs})$ indirectly as the stationary distribution of a Markov chain for which each transition step can be simulated directly. In particular, if we can simulate missing data under an assumed parameter value $\theta^{(t)}$,

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis}\,|\,Y_{obs},\theta^{(t)}\right), \qquad (2)$$

and simulate a new parameter value under a complete-data posterior that takes $Y_{mis} = Y_{mis}^{(t+1)}$,

$$\theta^{(t+1)} \sim P\left(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}\right), \tag{3}$$

then alternately performing (2) and (3) beginning from some starting value $\theta^{(0)}$ defines a Markov chain,

$$\left(Y_{mis}^{(1)}, \theta^{(1)}\right), \left(Y_{mis}^{(2)}, \theta^{(2)}\right), \ldots, \left(Y_{mis}^{(t)}, \theta^{(t)}\right), \ldots .$$

This algorithm is a special case of the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990), and it can be shown that, under very general conditions, the distribution of $\left(Y_{mis}^{(t)}, \theta^{(t)}\right)$ approaches the joint posterior distribution of Y_{mis} and θ, $P(Y_{mis}, \theta | Y_{obs})$, as $t \to \infty$. By taking t suitably large, $Y_{mis}^{(t)}$ will be, for all practical purposes, a draw from the desired distribution (1). Successive t*th* iterates $Y_{mis}^{(t)}, Y_{mis}^{(2t)}, Y_{mis}^{(3t)}, \ldots$ may be regarded as proper multiple imputations.

In order to sample from the complete-data posterior in (3), one needs to apply a prior distribution to θ. In practice, it is helpful to choose a prior from a natural conjugate family leading to a complete-data posterior that is easily simulated. When little is known about the parameters *a priori*, it may also be desirable to choose a prior distribution that is "noninformative" or relatively flat over the region of appreciable likelihood, so that Bayesian inferences will be nearly the same as inferences based only on the likelihood function. In our analysis, we chose the prior

$$p(\pi, \beta, \Sigma) \propto |\Sigma|^{-(q+1)/2} .$$

This prior is improper; it is not a true probability distribution because its integral over the parameter space is not finite. This prior assumes that the cell probabilities π_d are uniformly distributed over the unit interval subject to the constraint $\sum_{d=1}^{D} \pi_d = 1$, and that the regression coefficients in β are uniformly "distributed" over the entire real line (which is, of course, technically impossible). The exponent of -(q+1)/2 for the determinant of Σ was chosen by appealing to the Jeffreys invariance principle (e.g., Press 1982). Under this prior, the complete-data posterior distribution becomes the product of a Dirichlet distribution for π, an inverted Wishart distribution for Σ, and a matric-variate normal distribution for β given Σ, all of which are straightforward to simulate (Schafer, 1991).

4.2 COMPUTATIONAL PROCEDURES

We began our simulation by first finding $\hat{\theta}$, the ML estimate of θ, using the ECM algorithm of Little and Schlucter (1985). With the entire dataset stored in single precision and the parameters stored in double precision, model-fitting could be accomplished on a Sun SPARCstation ELC with 16 MB of main memory in approximately two hours. Under a very strict convergence criterion that required successive values of all parameters to change by less than 0.01%, the ECM algorithm converged in only 58 iterations.

After ML estimation, one imputation, which we shall call MI_0 was created under the assumption that $\theta = \hat{\theta}$. This initial set of imputed values cannot be regarded as one of the multiple imputations for purposes of inference because it fits the data "too well"—i.e., it does not reflect any uncertainty associated with the estimation of θ. Nevertheless, the MI_0 series was valuable in that it provided a set of typical imputes to examine the adequacy of the model.

Multiple imputations were generated by the iterative simulation scheme described above. A single Markov chain was started at $\theta^{(0)} = \hat{\theta}$ and allowed to run for 400 iterations. Every 40th value of Y_{mis} in the Markov chain was taken to be an independent draw from the stationary distribution. In this way, ten sets of imputations, which we shall call MI_1, MI_2,···, MI_{10}, were produced. The entire simulation took approximately 30 hours on a dedicated SPARCstation ELC.

The sequence MI_1, MI_2,···, MI_{10} can be considered proper multiple imputations only if the Markov chain achieves approximate stationarity (independence of the starting value) by 40 steps. Convergence to stationarity is difficult to assess, especially because of the high dimensionality of Y_{mis} and θ, but we informally monitored convergence by inspecting time-series plots of a few selected scalar functions of the parameter. Plots of marginal means of the "continuous" variables, given by

$$E(Z_j|\theta) = \sum_{d=1}^{D} r_d \mu_{dj}$$

for j=1,2,···,23, and a plot of the loglikelihood function, are shown in Figure 2. These plots show little long-range dependence, suggesting that for most purposes, the algorithm probably "converges" well within 40 iterations. The one notable exception is the mean of AE7; this item was 62.4% missing due to a skip pattern in the questionnaire. The long-range dependence in this plot suggests that the ten imputations for AE7 are somewhat correlated and could understate the missing-data uncertainty for this variable. AE7 is not a variable of primary interest in this study, but it is somewhat correlated with total serum cholesterol LGTCRES (in Table 5, the partial correlation is -.35), which is of primary interest. The lack of stochastic convergence with respect to some aspects of the distribution of AE7 could mean that the multiple imputations for total serum cholesterol are slightly correlated, and that multiple-imputation inference for this variable based on MI_1, MI_2,···, MI_{10} could understate the actual uncertainty.

5. EXPLORATORY ANALYSES OF THE IMPUTED DATA

5.1 GRAPHICAL DISPLAYS

This section presents some graphical displays and exploratory analyses of the imputed values. These are not intended to be a comprehensive evaluation; rather, the displays and discussion here are meant to be merely representative of the evaluations one could perform, and it is hoped that they convey some of the essential features of the imputation method.

Rather than examining all ten sets of multiple imputations, which would have been very tedious, we focused our attention on set MI_0. Set MI_0, although it is not one of the sets intended for use in multiple-imputation inference, is a natural choice for diagnostic analyses. MI_0 was produced under the ML estimate for θ, whereas MI_1-MI_{10} incorporate variability of the parameters about the ML estimate. Consequently, MI_0 represents the "best" fit of the model to the observed data; if examination of MI_0 reveals serious lack of fit, then the multiple imputations MI_1-MI_{10} should look even worse. On the other hand, any discrepancies between MI_0 and MI_1-MI_{10} might be due not to failings of the model, but merely to sampling variability.

Figure 3 displays histograms of the observed data, along with histograms of the imputed values in MI_0, for the 23 variables in the dataset subject to nonresponse. These 23 variables were all modeled as continuous even though seven of them are actually categorical; after imputation, the continuous imputes for these seven categorical variables were rounded off to the nearest categories. The general agreement between the marginal distributions of the observed and imputed values for most variables is quite striking. Since the model under which we are imputing assumes only that the missing data are missing at random rather than the more restrictive missing completely at random, the fact that the marginal distributions of the imputed values do not precisely match the observed marginals is not necessarily evidence of model failure, but could be due to the fact that the cases with missing values differ systematically from the rest of the sample on their observed characteristics. Surprisingly, though, Figure 3 shows that marginally, the

imputed values do match the observed values quite well for most variables, suggesting that the missing data may be approximately missing completely at random.

For interview variable AF10 (heart attack ever diagnosed?), the proportion of 1's in the imputed data is smaller than the proportion of 1's in the observed data, which could indicate that the normal model is not providing a good fit to this highly skewed dichotomous variable. For some of the blood pressure readings (like LGASYS), the imputed values are notably less skewed than the observed values, which is not surprising given the normal model. The histogram for observed values of LGAHT has an unusual shape because self-reported heights were recorded to the nearest inch; rounding and heaping gives this histogram a multimodal appearance, whereas the imputed values of LGAHT have a more normal appearance.

A good imputation method should accurately preserve not only the marginal distributions of the variables involved, but the relationships between variables as well. Selected scatterplots of pairs of MEC variables are displayed in Figure 4, separating the cases into those that had both variables observed and those that had either or both imputed. Figure 4 (a) plots the following pairs of body measurements: AHT versus HT, AWT versus WT, and WST versus BUT; all variables are plotted on the log scale, but axes are labeled in inches and pounds. Figure 4 (b) plots four pairs of systolic versus diastolic blood pressure readings (one from the interview, three from the MEC examination) with systolic blood pressures plotted on a log scale. Figure 4 (c) plots total serum cholesterol versus HDL cholesterol, both on the log scale. Inspection of 4 (a) reveals a small number of unusual outliers; these cases contained gross measurement errors which escaped the editing process described in Section 3.1. In each case, one of the recorded body measurements was highly erroneous, and the imputation procedure attempted to fill in the missing values in a manner consistent with those erroneous measurements. Except for these few unusual cases, the imputation model seems to have accurately preserved the correlations between these pairs of variables. Some of the more subtle non-normal features of the observed data, however, were not preserved in the imputed values--the trailing observations in the upper right-hand corner of the (WST, BUT) plot, for example.

Evidence of model failure for a particular variable could come not only by comparing the distributions of observed and imputed values across the whole sample, but within subclasses defined by other variables in the model. We generated a large number of plots of observed and imputed values within meaningful subgroups that were expected to be somewhat homogeneous with respect to the variables of interest, but that were still large enough to produce interpretable histograms and scatterplots. Quite generally, however, we found that the imputed values tended to mimic the observed values remarkably well, even within very fine subclasses of the population.

5.2 MULTIPLY-IMPUTED INTERVAL ESTIMATES

Using techniques described by Rubin (1987) for multiple-imputation inference about scalar estimands, we calculated standard errors and interval estimates for a number of quantities of interest based on the ten imputations MI_1-MI_{10}. Let Q denote a scalar quantity to be estimated. Let $\hat{Q}$ and U_i denote a complete-data point estimate and variance estimate for Q, respectively, calculated from the *ith* imputed dataset, i=1,2,...,m. In a typical survey setting, $\hat{Q}_i$ will be a weighted estimate (e.g., a Horvitz-Thompson estimate) and U_i will be calculated by some linearization or replication-based method that takes into account the complex sample design. When m multiple imputations are available, the natural point estimate for Q is simply

$$\bar{Q}_m = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i,$$

the average of the m complete-data point estimates.

The variance estimate associated with $\bar{Q}_m$ has two components. The *within-imputation* component, $\bar{U}_m$ is simply the average of the complete-data variance estimates,

$$\bar{U}_m = \frac{1}{m}\sum_{j=1}^{m} U_j,$$

and the *between-imputation* component, B_m, is the sample variance of the m complete-data point estimates,

$$B_m = \frac{1}{(m-1)}\sum_{j=1}^{m} (\hat{Q}_j - \bar{Q}_m)^2.$$

The *total variance*, T_m, is defined as

$$T_m = \bar{U}_m + (1 + m^{-1})B_m,$$

with the m^{-1} included for technical reasons. A 100(1-α)% interval estimate is formed by taking $\bar{Q}_m \pm t_\upsilon(t-\alpha/2)T_m^{1/2}$, where $t_\upsilon(p)$ denotes the 100p*th* percentile of the t distribution with υ degrees of freedom, and

$$\upsilon = (m-1)\left[1 + \frac{m\bar{U}_m}{(m+1)B_m}\right]^2.$$

The degrees of freedom υ lend insight into the efficiency of multiple imputation. A fully efficient interval estimate would be based on m=∞ imputations, would take $T_m = \bar{U}_m + B_m$, and would use the normal distribution rather than t_υ. Since t_υ approaches a standard normal as $\upsilon \to \infty$, a large value of υ suggests that the normal approximation would work well because the between-imputation component of variance B_m is well-estimated.

More insight into efficiency is provided by the it *relative increase in variance due to nonresponse*,

$$r_m = (1 + m^{-1})B_m/\bar{U}_m,$$

and the *fraction of missing information* due to nonresponse,

$$\gamma = \frac{r_m + 2/(\upsilon + 3)}{r_m + 1}$$

The value of r_m is the fraction by which a typical variance estimate from a singly-imputed dataset, approximated by $\bar{U}_m$, would need to be increased to account for missing-data uncertainty; and γ is an estimate of the proportion of total information (in the sense of Fisher) about Q in the complete data (Y_{obs}, Y_{mis}) that is contained in Y_{mis}. The total variance T_m is approximately proportional to $(1 + \gamma/m)$, which implies that unless γ is large, an inference based on a small number m of imputations is nearly as efficient as an inference based on m=∞ imputations. For example, if the fraction of missing information is 15%, then an inference based on m=3 imputations would have a total variance estimate of only about (1+.15/3)=1.05 times as great as T_∞. Hence, unless γ is large, there tends to be little advantage to using more than a small number of multiple imputations.

Multiple-imputation results for means of six MEC variables within categories of race/ethnicity are displayed in Table 6. The complete-data point estimates Q_i were calculated using basic survey weights (i.e., inverse probabilities of selection), without the adjustments for poststratification that will ultimately be incorporated. Complete-data variance estimates U_i were calculated with SUDAAN software (Shah et al., 1991) using linearization-based methods. Table 6 displays the within-imputation component of variance and the total variance on the standard deviation scale. For purposes of comparison, the table also displays Q_0 and U_0, the point and variance estimates from imputation set MI_0.

TABLE 6. Multiple-imputation results for means of six MEC variables within categories of race/ethnicity. All estimates are calculated using basic survey weights (inverse probabilities of section).

	Q_0	$U_0^{1/2}$	$\bar{Q}_{10}$	$\bar{U}_{10}^{1/2}$	$T_{10}^{1/2}$	$100r_{10}$	υ	100γ
Height (cm)								
White/other	168.33	0.207	168.25	0.201	0.206	5.7	3041	5.5
Black	168.22	0.182	168.22	0.192	0.200	8.2	1573	7.7
Mex-Amer	163.01	0.182	163.04	0.195	0.205	11.0	917	10.1
Weight (kg)								
White/other	73.91	0.388	73.85	0.364	0.380	9.2	1258	8.6
Black	77.62	0.494	77.58	0.506	0.542	14.8	543	13.2
Mex-Amer	71.98	0.481	72.20	0.451	0.465	6.3	2610	5.9
Avg.systolic BP								
White/other	121.33	0.517	121.46	0.487	0.515	11.8	814	10.7
Black	124.25	0.717	124.18	0.709	0.766	16.6	446	14.6
Mex-Amer	118.14	0.678	117.94	0.647	0.675	8.8	1368	8.2
Avg. diastolic BP								
White/other	72.75	0.382	72.66	0.395	0.416	10.8	938	10.0
Black	74.48	0.462	74.63	0.473	0.503	13.2	663	11.9
Mex-Amer	71.13	0.526	71.17	0.519	0.536	6.4	2500	6.1
Total cholesterol								
White/other	205.88	0.962	206.04	1.036	1.118	16.3	459	14.4
Black	200.76	1.056	201.28	0.960	1.075	25.3	221	20.9
Mex-Amer	200.16	2.402	200.28	2.447	2.490	3.6	7526	3.5
HDL cholesterol								
White/other	50.97	0.337	51.00	0.388	0.414	14.1	588	12.7
Black	55.95	0.510	55.95	0.462	0.514	23.7	244	19.8
Mex-Amer	50.07	0.601	50.03	0.505	0.524	7.6	1827	7.1

One striking feature of Table 6 is the large values of υ, which suggest that the between-imputation components of variance tend to be very well estimated. In contrast, we have good reason to suspect that the within-imputation components of variance are estimated rather poorly. Design effects (not shown) provided by SUDAAN displayed erratic behavior across the ten sets of multiple imputations (Little and Rubin, 1992). For example, the design effects for mean HDL cholesterol among Mexican-Americans across MI_1-MI_{10} ranged from 1.36 to 2.28, even though the low fraction of missing information (7.6%) for this quantity indicates that essential features of the sample data for this variable were changing very little across the imputation sets. The great variability we observed in the U_i's suggests that the methods of design-based variance estimation currently used for NHANES are inherently unstable, due perhaps to the small number of primary sampling units. Keeping this in mind, we interpret the results in Table 6 only with caution.

Another striking feature of this table is that the fractions of missing information γ are quite small, ranging from 20.9% down to 3.5%, even though the MEC variables in this dataset were missing at rates in excess of 30%. The fact that the fraction of missing *information* is often dramatically lower than the fraction of missing *observations* indicates that the imputation model was very effective in gleaning information about the missing data from the observed data. Comparing these values of γ with the multiple correlation coefficients R^2 from the exploratory regressions (Table 3), we see that γ tends to be lowest for the MEC variables that can be predicted from other variables with the greatest relative precision. This suggests that the gains in precision from a good imputation model, which makes intelligent use of information about Y_{mis} available in Y_{obs}, can be substantial.

The relative increases in variance due to nonresponse r_m in Table 6 range from 4 to 25%, so the total variances T_m tend to be not much larger than the within component $\bar{U}_m$. Hence, the multiply-imputed interval estimates for these quantities, at least, are not much wider than single-imputation intervals. Within smaller subdomains, however, we found that the values of r_m could be much higher. In particular, we found that the fractions of missing information for mean cholesterol (HDL and total) within subdomains of age, sex, and race were as high as 60%, and the intervals based on T_m were up to 60% wider than intervals based on $\bar{U}_m$.

For most uses of this dataset that we can imagine, it appears that m=10 imputations are more than enough to permit accurate and efficient inferences (Little and Rubin, 1992). With most fractions of missing information in the range 5-15%, it seems that m=5 or even m=3 would be adequate.

6. DISCUSSION

6.1 EFFICIENCY OF ESTIMATION

A comprehensive evaluation of this model-based multiple imputation procedure, and a comparison with the current NHANES weighting-class adjustments, is beyond the scope of this paper. Based on the evidence of Tables 1-6, however, it appears that for the twelve MEC variables in our study, we have achieved some substantial improvements in efficiency. The current weighting-class adjustments ignore much of the useful information in Y_{obs} relevant to predicting Y_{mis}. Our imputation method, however, was quite effective in using the information in Y_{obs}, as evidenced by the large discrepancies between the fractions of missing observations near 30% (Table 1) and the fractions of missing information near 15% (Table 6). By making intelligent predictions from Y_{obs}, the imputation model is "recovering" up to half of the information in the missing data.

Judging from the histograms and plots in Figures 3-4, the missing data in this dataset do not appear to be far from missing completely at random. This suggests that nonresponse bias may not be a serious problem, at least not for the twelve MEC variables in this study. The demographic and geographic variables used to define weighting classes have some predictive power (Table 5), but the weighting class adjustment ignores the relevant information available from the household interview, which can be substantial. Imputation, therefore, appears to offer meaningful improvements over the weighting-class adjustment in the reduction of variance.

6.2 VARIANCE ESTIMATION AND MISSING-DATA UNCERTAINITY

Once multiple imputations have been generated, it becomes a relatively simple matter to assess missing-data uncertainty for virtually any estimand of interest. The high values of υ in Table 6 suggest that our estimates of the between-imputation component of variance are quite good.

Although the current NHANES weighting-class adjustments make no provisions for assessing missing-data variance, two alternative proposals have been made in this regard. One proposal (Judkins and Winglee, 1992) involves switching the current variance estimation method from linearization, which is used by SUDAAN, to a replication-based procedure like the jackknife. With a replication-based procedure, it may be possible to measure missing-data uncertainty by calculating separate nonresponse-adjusted weights for each replicate. As discussed in Section 5.2, however, we suspect that in many cases the within-imputation or complete-data component of variance is being poorly estimated due to the small number of degrees of freedom. Others (see Longford, 1992 and his references) have found that replication methods like the jackknife may be inefficient and have poor sampling properties when the degrees of freedom are as small as they are here. To attempt to measure missing-data variance by this method, in addition to sampling variance, would almost certainly be less efficient than multiple imputation.

A second proposal for measuring missing-data uncertainty involves using imputation together with a variance multiplier of the form

$$A = \frac{n_r + n_n}{n_r + R^2 n_n} ,$$

where n_r is the number of respondents, n_n is the number of nonrespondents, and R^2 is the multiple squared correlation from the imputation model (Madow, Nisselson, and Olkin, 1983; Judkins and Winglee, 1992). The numerator of this multiplier is the apparent sample size after imputation, and the denominator is an effective sample size based on the predictive power of the imputation model. This type of adjustment has some major disadvantages relative to multiple imputation. First, it appears to be limited to the estimation of simple population means and proportions. Second, it is biased downward because it does not reflect any uncertainty associated with the fitting of regression model; the R^2 from the *estimated* regression will always be larger than the R^2 from the *true* regression. Finally, this type of multiplier implicitly assumes that the same multiple regression R^2 should apply to all nonresponding units, i.e., that the variable in question can be predicted with equal precision for all cases. This is clearly not the case. Consider two persons with missing values of weight--one with height, waist and buttocks circumference all recorded, and the other with no body measurements recorded. The R^2 ought to be much higher for the former. There is no simple way to address the varying degrees of information available for predicting missing values on a case-by-case basis, except a fully multivariate, multiple imputation procedure like the one we implemented.

6.3 THE VALIDITY OF OUR IMPUTATION MODEL

Fay (1992) and others have pointed out, quite correctly, that the validity of multiple-imputation inference depends on how the imputations are generated; if important aspects of the data or sample design are not included in the imputation model, anomalies may result. We believe that our imputation model does make intelligent use of the observed data Y_{obs}, including the pertinent information on sample design, by using dummy indicators for STAND and AGE × SEX × RACE as predictors for each of the 23 variables subject to nonresponse.

A purist might argue that to ensure validity of the multiple imputations from a design perspective, one ought to fit a separate imputation model within each primary sampling unit (PSU), or perhaps include in the imputation model more interactions involving STAND. We actually considered using this approach until it became apparent that the data within PSU's was too sparse to fit a good model, a model that satisfactorily preserved many of the relationships that we considered as important or more important than the PSU effects. A more elaborate solution would have involved borrowing strength across PSU's through a hierarchical model, which corresponds to treating the PSU effects as random rather than fixed. Our model lacks the conceptual appeal of a hierarchical model, but it was computationally much easier to work with. Judging from Table 5, the main effect of STAND seems to account for at most $.28^2$=8% of the variance in the MEC variables. Our experience suggests that higher-order interactions involving STAND would be of a smaller magnitude than the main effects, and would account for an even smaller portion of the

variance. We do not believe that including more interactions, fitting separate models for each PSU, or a hierarchical approach would have given answers substantially different from what we obtained with our relatively simple model.

As discussed in Section 5.2, we believe that in this application, the between-imputation components of variance are far better estimated that the within-imputa\-tion (i.e., complete-data) components of variance. If the multiple-imputation variance estimates of Table 6 are not to be believed, it is not because the imputation model and simulation procedures are flawed, but because SUDAAN and the classical design-based variance estimation methodology is inadequate in this setting. Strengthening the techniques of complete-data variance estimation, in addition to properly reflecting missing-data uncertainty, is also an important priority (Little and Rubin, 1992).

6.4 COMPUTATIONAL CONSIDERATIONS

The advantages of model-based multiple imputation over the current NHANES weighting-class methodology must be weighed against increased computational demands. Weighting-class adjustments, at least as they are carried out now, do not involve explicit probability modeling of the data; no model fitting or simulation is needed; and nonresponse adjustments for all variables are made at once through a single set of weights. Computationally, then, weighting-class adjustments are much simpler to implement. On the other hand, weighting-class adjustments do not address the problem of item nonresponse, which is a small but non-negligible part of the NHANES missing-data problem. Ad hoc methods for dealing with item nonresponse, such as hot-deck imputation, are complex and difficult to implement in multivariate datasets with complex patterns of missingness. Our method of model-based imputation solves the problems of both unit and item nonresponse simultaneously.

6.5 THE IMPORTANCE OF THIS WORK

This work, to our knowledge, is the first fully Bayesian application of multiple imputation in a multivariate survey setting, where multiple imputations are drawn from a posterior predictive distribution for Y_{mis} given Y_{obs} under an explicit model. Previously, Kennickell (1991) has applied Markov-chain simulation to create multiple imputations for over 200 variables in a large multivariate database. His work bears some important similarities to ours, but it does not include an explicit specification for the full joint probability distribution of the complete data.

We believe this work represents an important advance in the practice of survey nonresponse adjustments. Using state-of-the-art techniques, we were able to multiply impute a dataset with 27 variables and 12,392 observations without much difficulty. It does not appear possible at present to multiply impute the entire NHANES survey; at least it is not possible to impute all variables at once under a general multivariate model. It is evident, however, that newly developed techniques of simulation, and our ever-increasing computational power, are providing an excellent set of tools for handling incomplete survey data, and that our capabilities in this area will continue to increase in the years ahead.

REFERENCES

BECKER, R.A., CHAMBERS, J.M., and WILKS, A.R. (1988), *The New S Language*, Pacific Grove, CA: Wadsworth and Brooks/Cole Advanced Books and Software.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

COX, B. (1991), "Weighting survey data for analysis," presentation for the ASA Continuing Education Program, Joint Statistical Meetings, August, 1991.

EZZATI, T. and KHARE, M. (1991), "Consideration of health variables to adjust sampling weights for nonresponse in a national health survey," *Proceedings of the Social Statistics Section of the American Statistical Association*, 203-208.

EZZATI, T. and KHARE, M. (1992), "Nonresponse adjustments in a national health survey," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, in press.

EZZATI, T., MASSEY, J., WAKSBERG, J,, CHU, A., and MAURER, K. (1992), "Sample design: Third National Health and Nutrition Examination Survey," National Center for Health Statistics, Vital Health Statistics, Series 2, No. 113.

FAY, R.E. (1992), "When are inferences from multiple imputation valid?" *Proceedings of the Survey Research Methods Section of the American Statistical Association*, in press.

GELFAND, A.E. and SMITH, A.F.M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398-409.

GEMAN, S. and GEMAN, A. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-740.

JUDKINS, D. and WINGLEE, M. (1992), "Variance estimation with imputed data for NHANES III," memorandum dated 9/17/92, Rockville, MD: Westat.

KENNICKELL, A.B. (1991), "Imputation of the 1989 Survey of Consumer Finances: stochastic relaxation and multiple imputation," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1-10.

KRZANOWSKI, W.J. (1980), "Mixtures of continuous and categorical variables in discriminant analysis," *Biometrics*, 36, 493-499.

KRZANOWSKI, W.J. (1982), "Mixtures of continuous and categorical variables in discriminant analysis: a hypothesis testing approach," *Biometrics*, 38, 991-1002.

LITTLE, R.J.A. (1986), "Survey nonresponse adjustments for estimates of means," *International Statistical Review*, 54, 139-157.

LITTLE, R.J.A. and RUBIN, D.B. (1992), "Assessment of Trial Imputations of NHANES: Final Report," Waban, MA: Datametrics Research.

LITTLE, R.J.A., and SCHLUCTER, M.D. (1985), "Maximum-likelihood estimation for mixed continuous and categorical data with missing values," *Biometrika*, 72, 492-512.

LONGFORD, N.T. (1992), "Comparison of efficiency of jackknife and variance component estimators of standard errors," Program Statistics Research Technical Report 92-24, Educational Testing Service, Princeton, NJ.

MADOW, W.G., NISSELSON, H., and OLKIN, I., eds.\(1983), *Incomplete Data in Sample Surveys*, Volume 1, New York: Academic Press.

MADOW, W.G., OLKIN, I., and RUBIN, D.B., eds. (1983), *Incomplete Data in Sample Surveys*, Volume 2, New York: Academic Press.

MULLER, P. (1991), "A generic approach to posterior integration and Gibbs sampling," Technical Report 91-09, Department of Statistics, Purdue University, West Lafayette, IN.

MENG, X.L., and RUBIN, D.B. (1992), "Maximum-likelihood estimation via the ECM algorithm: a general framework," to appear in *Biometrika*.

PRESS, S.J. (1982), *Applied Multivariate Analysis* (second edition), New York: Holt, Rinehart, and Winston.

RUBIN, D.B. (1976), "Inference and missing data," *Biometrika*, 63, 581-592.

RUBIN, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley \& Sons.

SCHAFER, J.L. (1991), "Algorithms for multiple imputation and posterior simulation from incomplete multivariate data with ignorable nonresponse," Ph.D. Dissertation, Department of Statistics, Harvard University, Cambridge, MA.

SCHAFER, J.L. (1993), *Analysis of Incomplete Multivariate Data by Simulation*, New York: Chapman and Hall, in preparation.

SHAH, B.V., BARNWELL, B.G., HUNT, P.N., LAVANGE, L.M. (1991), *SUDAAN User's Manual: Professional Software for Survey Data Analysis for Multistage Sample Designs, Release 5.0*, North Carolina: Research Triangle Park.

FIGURE 2. Time-series plots of marginal means of 23 "continuous" variables of 400 iterations of the Markov chain simulation. Shown also is a time-series plot of the loglikelihood function.

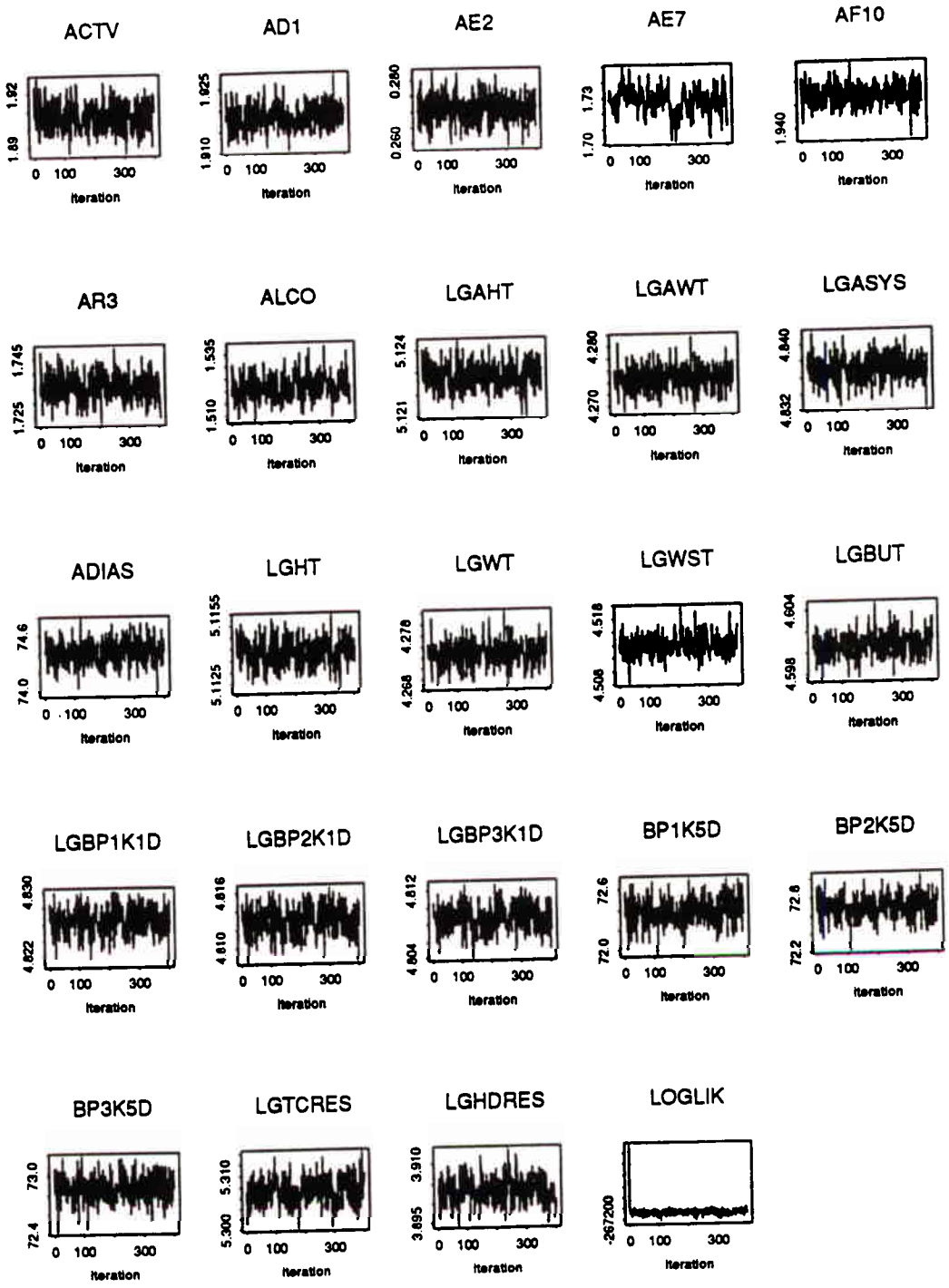


FIGURE 3. Marginal histograms of observed values and imputed values generated under the imputation model, set MI_0

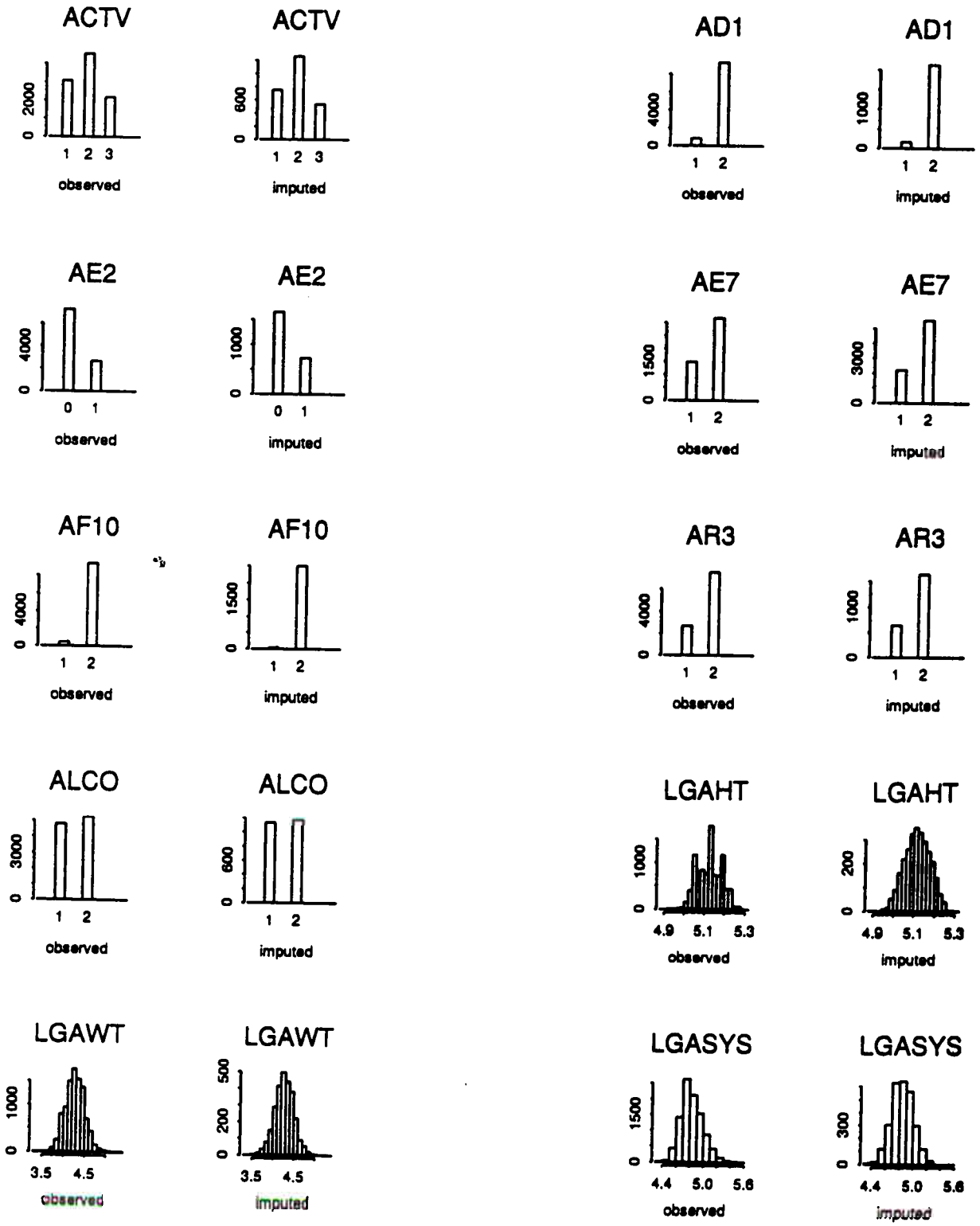


FIGURE 3. Marginal histograms of observed values and imputed values generated under the imputation model, set (continued)

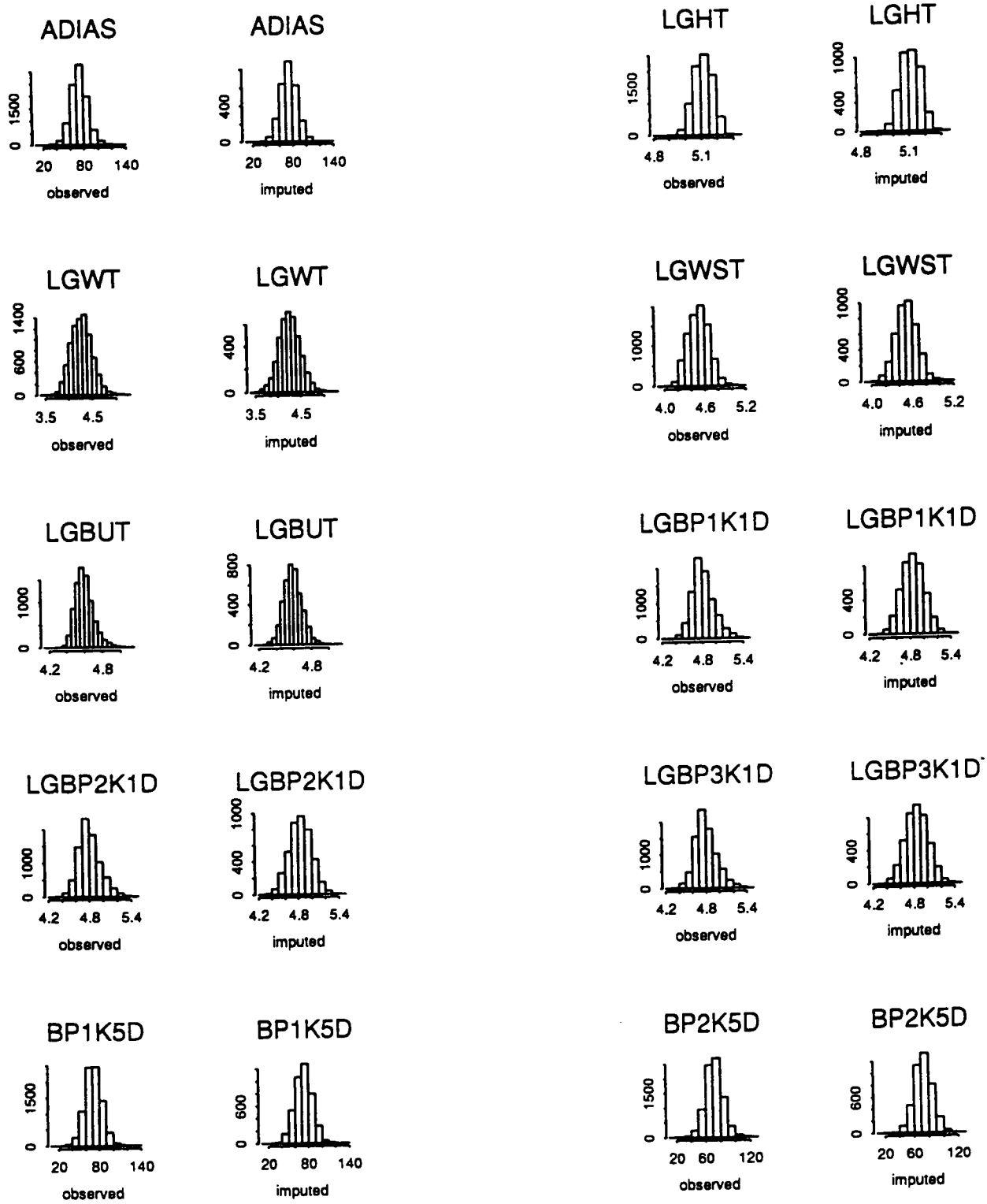


FIGURE 3. Marginal histograms of observed values and imputed values generated under the imputation model, set (continued)

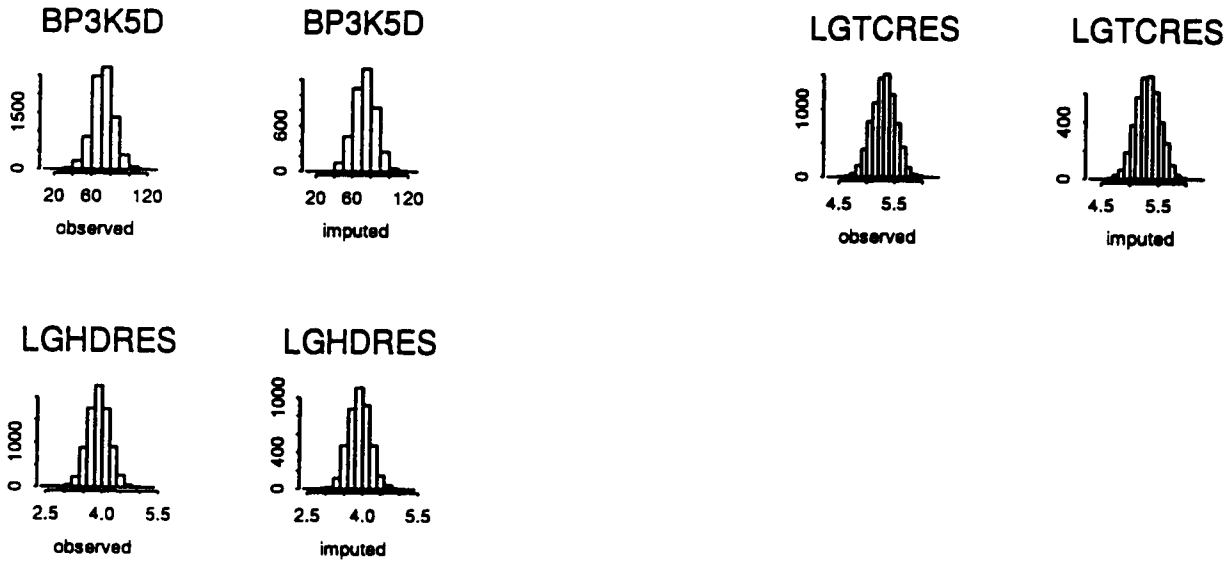


FIGURE 4a. Selected bivariate scatterplots of body measurements for cases having both variables observed, and cases having one or both variables imputed, set MI_6

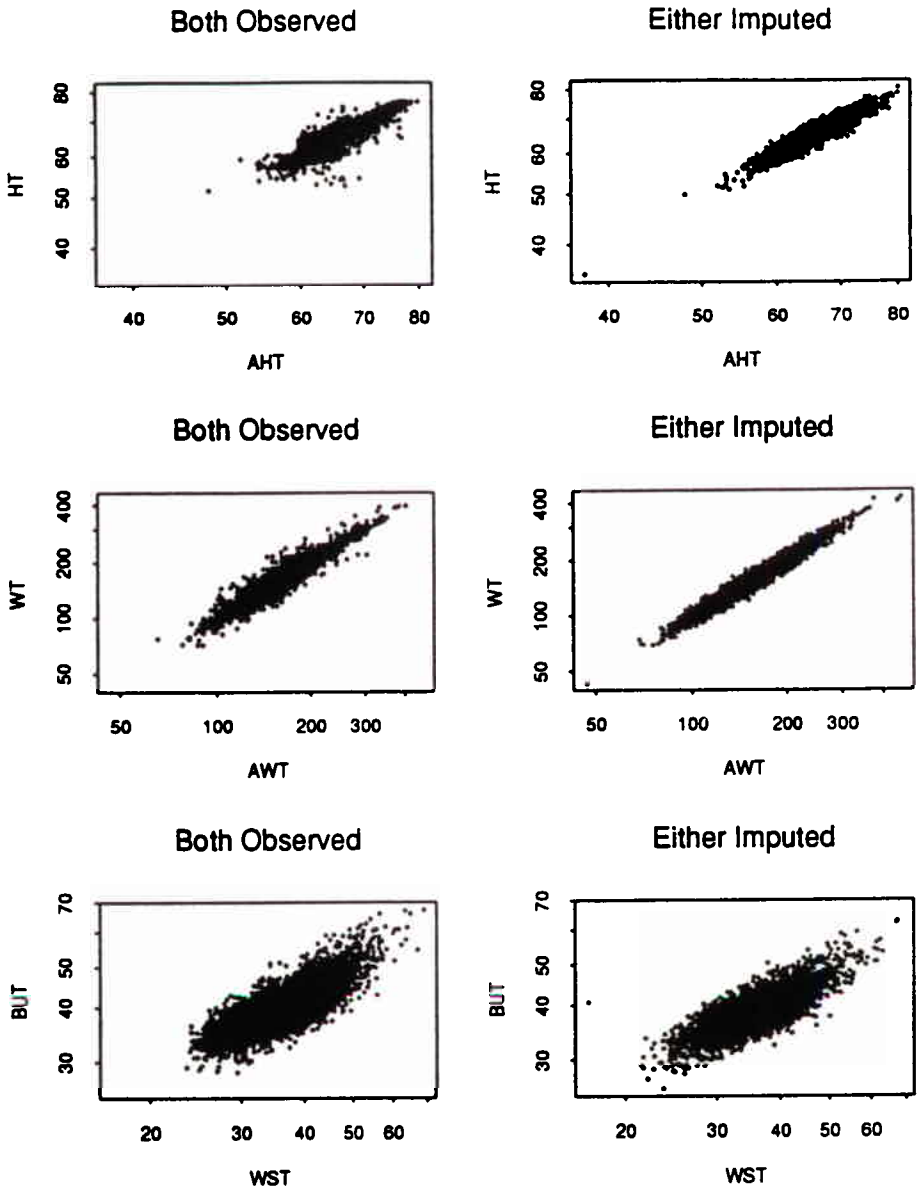


FIGURE 4b. Bivariate scatterplots of systolic versus diastolic blood pressure for cases having both variables observed, and cases having one or both variables imputed, set MI_0

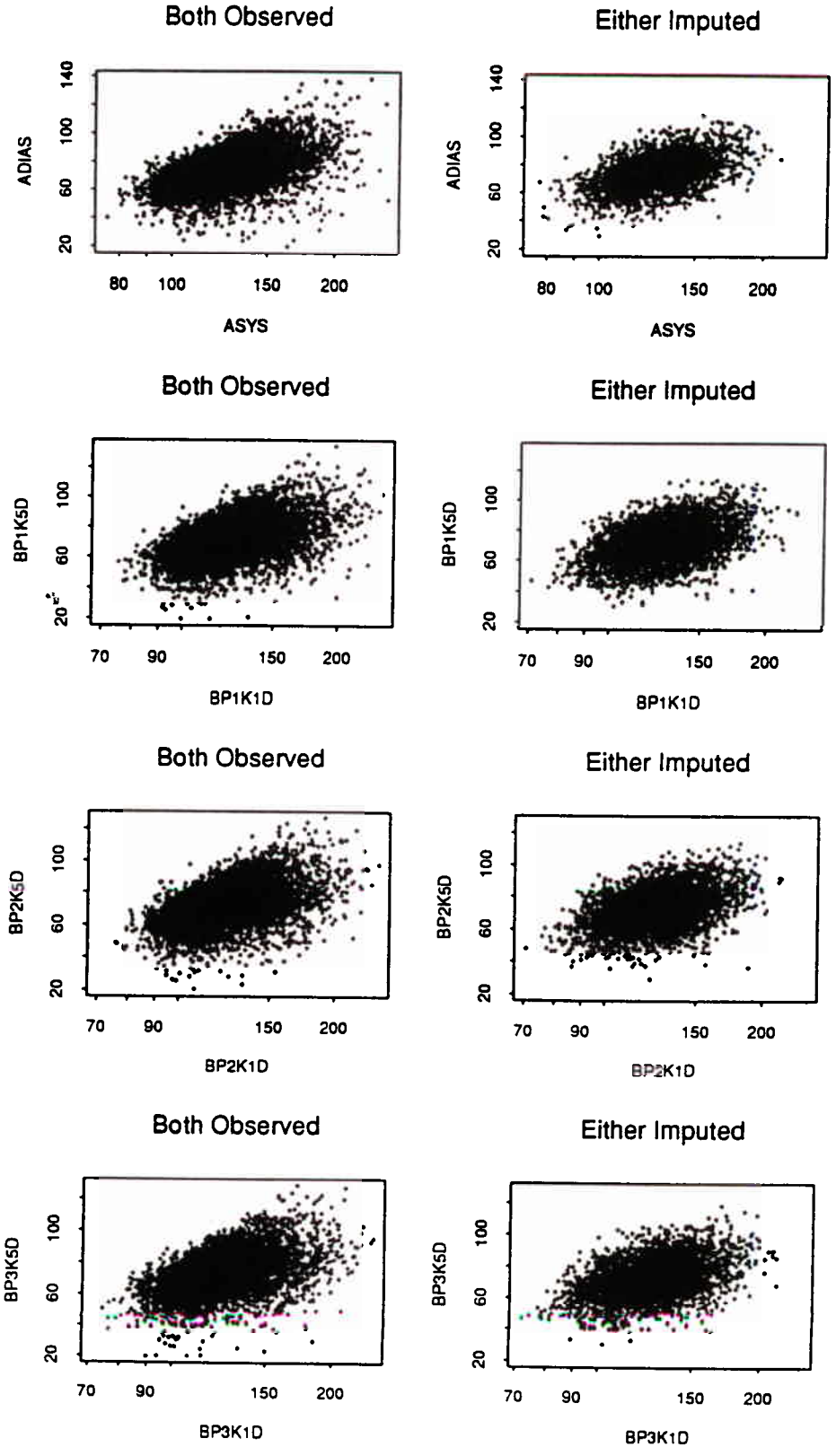


FIGURE 4c. Bivariate scatterplots of total serum cholesterol versus HDL cholesterol for cases having both variable observed, and cases having one or both variables imputed, set MI_0